Rev. Mexicana Astron. Astrof., **21**, 674-676 (1990)

HOMOGENEITY AND ENTROPY

H. Tignanelli, R.A. Vazquez, C. Mostaccio
S. Gordillo and A. Plastino

Observatorio Astronómico
Universidad Nacional de La Plata

RESUMEN. Presentamos una metodología de análisis de la homogeneidad a partir de la Teoría de la Información, aplicable a muestras de datos observacionales.

ABSTRACT: Standard concepts that underlie Information Theory are employed in order design a methodology that enables one to analyze the homogeneity of a given data sample.

Key words: DATA ANALYSIS

I. INTRODUCTION

It is one of the main goals of statistical analysis to infer messages employing an appropriate methodology that carry information concerning the data set one is working with. One often needs to a-priory ascertain that this data sample exhibits reasonable characteristic of coherence that render it suitable for the purposes one has in mind. One speaks, within this context, of the homogeneity of the data sample, which is an a-priori measure of the quality of the data one is dealing with. Each element of this sample makes its individual contribution to the overall picture, that is, the homogeneity of the sample should be the result of some appropriate sum of individual characteristics that render the complete set useful for a certai purpose.

Our main idea is to associate to each data sample a probability distribution (and to each of its members a probability element) that is correllated with its degree of internal coherence, or homogeneity. For an ideally homogeneous sample this distribution is the uniform one, and it is seen that each element "contributes" equally. Any element is a faithful representative of the set.

Starting with this ideal situation we can think of associating a probability element to each of the members of the sample, so that we obtain a probability distribution (p.d.) for the set. The closer the resemblance between this p.d. and the uniform one, the more homogeneous our data sample is. We give below a more mathematical criterium to deal with the concepts here outlined.

II. AN EXAMPLE

Let us discuss here an specific example, that allows us to give concrete meaning to the considerations of the preceding Section. Suppose one is interested in star evolution theories constructed on the basis of UBURI data concerning open clusters. Assume N clusters are involved. Some questions immediately arise, concerning the concomitant data:

1) How large a distortion arises as a consequence of mixing photoelectric and photografic photometries?

2) What is the effect upon the quality of our data sample of the internal and external errors?

3) Suppose that for a given cluster just a single set of observational data is available.

ystematic departures from the standard system caused by this sample affect the entire sample all clusters). In which way?

) The fact that observations made by different authors are to be employed generates a ertain amount of distortion. How large?

The above questions, and related ones obviously affect the homogeneity of a given ata sample. In order to proceed according to our methodology, we consider for our N-cluster ata the following attributes:

) Internal photometric error
) External photometric error
) Type of photometry
) Magnitude range
) Number of star in each cluster
) Number (n) of different authors

Let "i" label each cluster. Our main idea is that of assigning a probability P_i to he homogeneity (or, more precisely, lack of it) of the sample under study. To the N clusters e thus associate a probability distribution. For an ideally homogeneous sample we have

$$P_1 = P_2 = \ldots = P_i = \ldots = P_N$$

Information theory (Duering et al, 1985a) provides a natural measure of the lack f homogeneity by recourse to the concept of entropys, which is naturally associated to any robability distribution (Duering et al., 1985a)

$$S = -C \sum_{i=1}^{N} P_i \ln P_i \ ,$$

here C is the constant that measures our information unit (Duering et al., 1985a). For the deally homogeneous sample S is an absolute maximum (Duering et al. 1985a)

$$S_{ideal} = C \ln N$$

o that $S_{ideal} - S > 0$ measures the"data distortion" of our sample. The task one faces is thus hat of evaluating S. This in turn implies having at our disposal a systematic procedure to ssign the values P_i on the basis of the available data.

II. GENERAL FORMALISM

We return now to the general, abstract situation, that must be specialized to the haracteristics of each particular astronomical problem.

We assume that our sample consists of N elements labelled by the subindex "i"

$$S_i = - C \sum_{i=1}^{N} P_i \ln P_i \qquad (1)$$

However, for each "i" we consider η pieces of data $f_i^{(k)}$, k = 1,2,..., η. Moreover, e suppose that the sum

$$\sum_{i=1}^{N} P_i f_i^{(k)} = f_k \ ; \ k = 1, \ldots, \eta \ , \qquad (2)$$

xists, and is both well-defined and available. The f_k constitute the essential ingredient hat enables one to determine the P_i , via the so-called Maximum Entropy Principle (MEP). The

idea is to extremalize the expression (1) with the constraints (2). Normalization provides an additional constraint

$$\sum_{i=1}^{N} P_i = 1 \qquad (3)$$

to the η ones arrising from (2). According to standard variational theory N + 1 Lagrange multipliers λ_0 , λ_1 , ..., λ_η will do the job (extremalizing S (Duering et al., 1985a)).

Let us remark that the f_k constitute actual pieces of data, that are re-interpreted as crising out of the particular composition of the N individual contributions given by (2). The variational problem (Duering et al., 1985a)

$$\delta \{ S - \lambda_o \sum_{i=1}^{N} P_i - \sum_{k=1}^{\eta} \lambda_k \sum_{i=1}^{N} P_i f_i^{(k)} \} = 0 \qquad (4)$$

has an exact, analytical solution (Duering et al., 1985a)

$$P_i = \exp \{ - \lambda_o - \sum_{k=1}^{\eta} \lambda_k f_i^{(k)} \} , \qquad (5)$$

where

$$\lambda_o = - \ln \sum_{i=1}^{N} \exp \{ \sum_{k=1}^{\eta} \lambda_k f_i^{(k)} \} \qquad (6)$$

are obtained by solving the coupled system (Duering et al., 1985b)

$$f_k = \frac{\partial \lambda_o}{\partial \lambda_k} \qquad (7)$$

Finally, one can rewrite the entropy as a function of the Lagrange multipliers (Otero et al., 1982)

$$S = C \lambda_o + C \sum_{k=1}^{\eta} \lambda_k f_k \qquad (8)$$

IV. SUMMARY

Summing up, we propose a methodology which associates a definite real number, the entropy S, with the homogeneitvof the system . We thus have at our disposal a quantitative measure of the up to now more or less vaguely defined idea of homogeneity.

REFERENCES

Duering, E., Otero, D., Plastino, A., Proto, A. 1985a, Phys.Rev., A32, 2455.
Duering, E , Otero, D., Plastino, A., Proto, A. 1985b, Phys.Rev., A32, 3681.
Otero, D., Plastino, A., Proto, A., Zannoli, G., 1982, Phys.Rev., A26, 1209.

ACKNOWLEDGMENTS

R.A. Vazquez and A. Plastino acknowledge support from CONICET Argentina, H.Tignanelli, C. Mostaccio and S. Gordillo thank the CIC, Provincia de Buenos Aires, for its support. H. Tignanelli and R. Vazquez acknowledge specially to IAU for their kind hospitality during the 6th Regional IAU Meeting, Gramado (Brasil), October 1989.

Horacio Tignanelli: Observatorio Astronómico, Paseo del Bosque S/N, (1900) La Plata,Argentina.